Impact of time and note duration tokenizations on deep learning symbolic music modeling

Nathan Fradet, Nicolas Gutowski, Fabien Chhel, Jean-Pierre Briot

IMPACT OF TIME AND NOTE DURATION TOKENIZATIONS ON DEEP LEARNING SYMBOLIC MUSIC MODELING

Nathan Fradet[1,2] **Nicolas Gutowski**[3] **Fabien Chhel**[3,4] **Jean-Pierre Briot**[1]
[1] Sorbonne University, CNRS, LIP6, F-75005 Paris
[2] Aubay, Boulogne-Billancourt, France
[3] University of Angers, LERIA, 49000 Angers, France
[4] ESEO-TECH / ERIS, 49100 Angers, France
nathan.fradet@lip6.fr

ABSTRACT

Symbolic music is widely used in various deep learning tasks, including generation, transcription, synthesis, and Music Information Retrieval (MIR). It is mostly employed with discrete models like Transformers, which require music to be tokenized, i.e., formatted into sequences of distinct elements called tokens. Tokenization can be performed in different ways. As Transformers can struggle at reasoning, but capture more easily explicit information, it is important to study how the way the information is represented for such model impact their performances. In this work, we analyze the common tokenization methods and experiment with time and note duration representations. We compare the performances of these two impactful criteria on several tasks, including composer and emotion classification, music generation, and sequence representation learning. We demonstrate that explicit information leads to better results depending on the task.

1. INTRODUCTION

Most tasks involving using deep learning with symbolic music [1] are performed with discrete models, such as Transformers [2]. To use these models, the music must first be formatted into sequences of distinct elements, commonly called tokens. For instance, a token can represent a note attribute or a time event. The set of all known tokens is commonly called the vocabulary, and each token is associated to a unique integer id. These ids are used as input and output of models.

Compared to text, tokenizing music provides greater flexibility, as a musical piece can be played by different instruments and composed of multiple simultaneous notes, each having several attributes to represent. As a result, it is necessary to serialize these elements along the time dimension. To achieve this, researchers have developed various methods of tokenizing music [3–6].

While these works present model performance comparisons between tokenizations, their main differences or similarities are not always clearly stated. Moreover, they mostly focus on music generation, for which evaluations are performed on results obtained autoregressively, which accumulates biases [7] and is arguably difficult to evaluate [8], rather than music modeling more broadly. Yet, Transformer models struggle at reasoning, i.e. making logical deduction based on information points in the input data [9, 10], but perform tasks better when fed with explicit information and instructions [11]. In the case of symbolic music, it is thus important to study how the ways the music information is represented impact model performances.

This paper's primary contribution is a thorough and well-designed comparison of common tokenization techniques. Our focus is on two critical aspects: the representation of time and note duration. We believe that they are significant and impactful design choices for any music tokenization approach. Through experiments on composer classification, emotion classification, music generation, and sequence representation, we demonstrate that these design choices produce varying results depending on the task, model type, and inference process. Autoregressive generation benefits from explicit note duration and time shift tokens, while explicit note offset is more discriminating better suited for contrastive learning approaches.

We present next the related works, followed by an analysis of music tokenization, experimental results, and finally a conclusion. The source code is publicly available. [1]

2. DECOMPOSING MUSIC TOKENIZATION

2.1 Related works

Early works using discrete models for symbolic music, such as DeepBach [12] or FolkRNN [13], rely on specific tokenizations often tied to their training data. Since then, researchers introduced more general representations applicable to any kind of music. The most commonly used are *Midi-Like* [3] and *REMI* [4]. The former tokenizes music

[1] https://github.com/Natooz/music-modeling-time-duration

by representing tokens as the same types of events from the MIDI protocol, while the latter represents time with *Bar* and *Position* tokens and note durations with explicit *Duration* tokens. Additionally, *REMI* includes tokens with additional information such as chords and tempo.

More recently, researchers have focused on improving the efficiency of models with new tokenizations techniques: *Compound Word* [14], *Octuple* [5] and *PopMAG* [15] merge embedding vectors before passing them to the model; 2) LakhNES [16] and [17], SymphonyNet [18] and [19] use tokens combining several values, such as pitch and vocabulary.

2.2 Music tokenization design

When analyzing the possible designs of music tokenization, we can distinguish seven key dimensions:

- **Time**: Type of token representing time, either *TimeShift* indicating time movements, or *Bar* and *Position* indicating new bars and the positions of the notes within them. We can also consider the unit of *Time-Shift* tokens, either in beats or in seconds. [2]
- **Notes duration**: How notes durations are represented, with either *Duration* or *NoteOff* tokens.
- **Pitch**: Most works use tokens representing absolute pitch values, although recent work shed light on the expressiveness gain of representing as intervals instead [20];
- **Multitrack representation**: The representation of several music tracks in a sequence, i.e., how are the notes linked to their associated track.
- **Additional information**: Any additional information such as chords, tempo, rests, note density. Velocity can also falls in this category;
- **Downsampling**: How "continuous-like" features are downsampled into discrete sets, e.g. the 128 velocity values reduced to 16 values;
- **Sequence compression**: Methods to reduce the sequence lengths, such as merging tokens and embedding vectors.

As time and note duration can both be represented in two different ways, existing tokenizations can be easily classified based on these dimensions, as shown in table 1. However, other dimensions offer a broader spectrum of potential designs.

For instance multitrack can be represented by `Program` tokens [3] preceding notes as in FIGARO [22], distinct tracks sequences separated by `Program` tokens as in MMM [23], combined note and instrument tokens as LakhNes [16] and MuseNet [17], or merging `Program` embeddings with the associated note tokens (MMT [24], MusicBert [5]). One could even infer each sequence separately and lately model their relationships with operations aggregating their hidden states an in ColBERT [25].

[2] In this paper we only treat of the beat unit. The MIDI protocol represents time in *tick* unit, which value is proportional to the time division (in ticks per beat) and tempo. Hence, working with seconds would require a conversion from ticks.

[3] Following the conventional programs from the MIDI protocol.

	Time		Note duration	
Tokenization	`TimeShift`	`Bar + Pos.`	`Duration`	`NoteOff`
MIDI-Like [3]	√	-	-	√
REMI [4]	-	√	√	-
Structured [21]	√	-	√	-
TSD [19]	√	-	√	-
Octuple [5]	-	√	√	-

Table 1: Time and note duration representations of common tokenizations. `Pos.` stands for Position.

The MIDI protocol supports a set of effects and metadata that can also be represented when tokenizing symbolic music, such as tempo, time signature, sustain pedal or control changes. Some works also include explicit `Chord` tokens, detected with rule-based methods. Nevertheless, only a few works experimented with such additional tokens so far ([4, 26]).

Previous works have mainly compared tokenization strategies by evaluating models with automatic and sometimes subjective (human) metrics, but often do not proceed to comparisons between the ways to represent one of the dimensions we introduced previously. [4] compared results for the generation task, for the use of `Bar` and `Position` tokens versus `TimeShift` in seconds and beats.

To the best of our knowledge, no comprehensive work and empirical analysis have fairly compared these possible tokenization choices. Conducting such an assessment would require an extensive survey. In this paper, we specifically focus on the time and note duration representations, as they are the two main characteristics present in every tokenization.

We want to highlight the importance of the explicit information carried by the token types, as they directly impact the performances of models. `TimeShift` tokens represent explicit time movements, and especially the time distances between successive notes. On the other hand, `Bar` and `Position` tokens bring explicit information on the absolute positions (within bars) of the notes, but not the onset distances between notes. One could assume that the former might help to model melodies, and the latter rhythm and structure. For note duration, `Duration` tokens intuitively express the absolute durations of the notes, while `NoteOff` tokens explicitly indicates the offset times. With `NoteOff`, a model would have to model note durations from the combinations of previous time tokens.

Our experiments aim to demonstrate the impact of different combinations of time and note duration tokens on model performance and which combinations are suitable for different tasks. Next, we introduce our methodology.

3. METHODOLOGY

3.1 Models and trainings

For all experiments, we use the Transformer architecture [2], with the same model dimensions: 12 layers, with dimension of 768 units, 12 attention heads and inner feed-forward layers of 3072.

For classification and sequence representation, it is first pretrained on 100k steps and a learning rate of 10^{-4}, then finetuned on 50k steps and a learning rate of 3×10^{-5}, with a batch size of 48 examples. An exception is made for the EMOPIA dataset, for which we set 30k pretraining steps and 15k finetuning steps, as it is fairly small. These models are based on the BERT [27] implementation of the Transformers library [28]. We use the same pretraining than the original BERT: 1) from 15% of the input tokens, 80% is masked with a special `MASK` token, and 20% is randomized; 2) half of the inputs have 50% of their tokens (starting from the end) shuffled and separated with a special `SEP` token, and the model is trained to detect if the second part is the next of the first.

For generation, the model is based on the GPT2 implementation of the Transformers library [28]: it uses a causal attention mask, so that for each element in the sequence, the model can only attend to the current and previous elements. The training is performed with teacher forcing, the cross-entropy loss is defined as: $\ell = -\sum_{t=1}^{n} \log p_\theta(x_t|\mathbf{x}_{\leq n})$.

All trainings are performed on V100 GPUs, using automatic mixed precision [29], the Adam optimizer [30] with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$, and dropout, weight decay and a gradient clip norm of respectively 10^{-1}, 10^{-2} and 3. Learning rates follow a warm-up schedule: they are initially set to 0, and increase to their default value during the first 30% of training, then slowly decrease back to 0.

10% of the data is used for validation during training, and 15% to test models. Inputs contains 384 to 512 tokens, and begin with a `BOS` (Beginning of Sequence) token and end with a `EOS` (End of Sequence) one.

3.2 Tokenizations

We investigate here the four combinations of possible time and note duration representation. In the results, we refer to them as *TS* (`TimeShift`), *Pos* (`Position`), *Dur* (`Duration`) and *NOff* (`NoteOff`). It is worth noting that *TS* + *Dur* is equivalent to *TSD* [19] and *Structured* [21], *TS* + *NOff* is equivalent to *MIDI-Like* [3], and *Pos* + *Dur* is equivalent to *REMI* (without additional tokens for chords and tempo).

We apply different resolutions for `Duration` and `TimeShift` token values: those up to one beat are downsampled to 8 samples per beat (spb), those from one to two beats to 4 spb, those from two to four beats to 2 spb, and those from four to eight beats to 1 spb. Thus, short notes are represented more precisely than longer ones. `Position` tokens are downsampled to 8 spb, resulting in 32 different tokens as we only consider the 4/* time signature. This allows to represent the 16^{th} note. We only consider pitches within the recommended range for piano (program 0) specified in the General MIDI 2 specifications[4]: 21 to 108. We then deduplicate all duplicated notes. Velocities are downsampled to 8 distinct values. No additional token (e.g., `Chord`, `Tempo`) is used.

We perform data augmentation by creating variations of the original data with pitches increased and decreased by two octaves, and velocity by one value. Finally, following [19], we use Byte Pair Encoding to build the vocabularies up to 2k tokens for generation and 5k for other tasks. All these preprocessing and tokenization steps were performed with MidiTok [6].

4. GENERATION

For the generative task, we use the POP909 dataset [31]. The models start with prompt made of between 384 to 512 tokens, then autoregressively generate 512 additional tokens. Evaluation of generated results remains an open issue [8]. Previous work often perform measures of similarity of certain features such as pitch range or class, between prompts and generated results, alongside human evaluations. Feature similarity is however arguably not very insightful: a generated result could have very similar features to its prompts while being of poor quality. Human evaluations, while being more reliable on the quality can also induce biases. Besides, [4] already shows results on an experiment similar to ours.

Hence we choose to evaluate results on the ratios of prediction errors: Token Syntax Error (TSE) [19]. This metric is bias-free and directly linked to the design choices of the tokenizations. It allows us to measure how a model achieves to make reliable predictions based on the input context and the knowledge it learned.

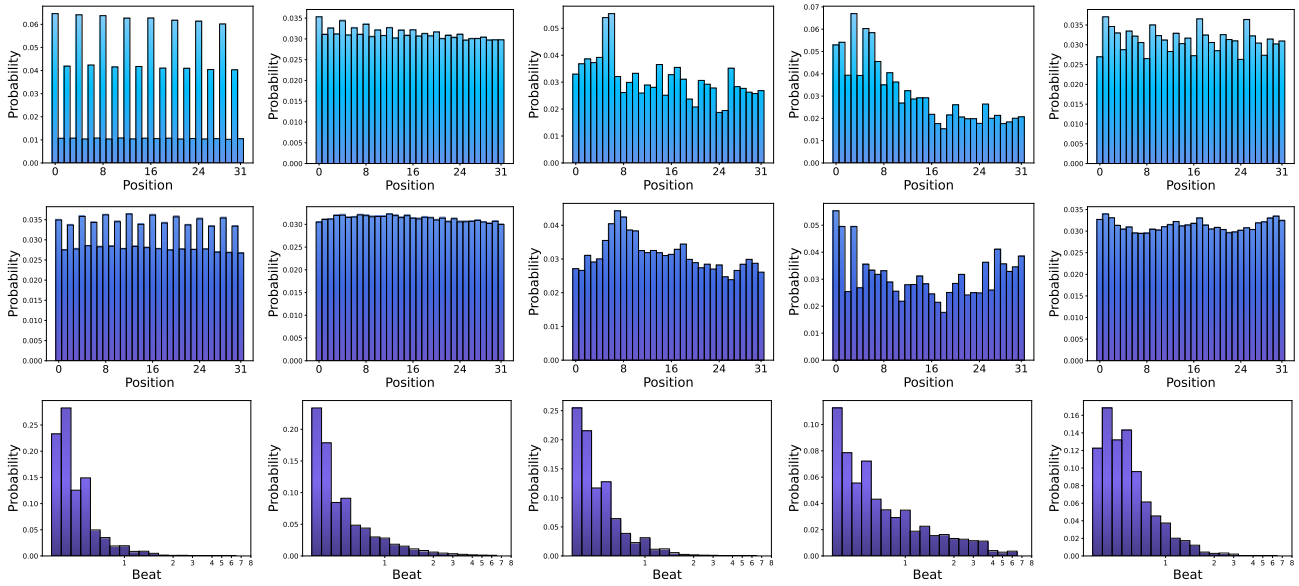
We use the categories from [19]:

- $\mathbf{TSE_{type}}$: an error of type, e.g., when the model predicts a token of an incompatible type with the previous one.
- $\mathbf{TSE_{time}}$: a wrong predicted `Position` value, that goes back or stay in time.
- $\mathbf{TSE_{dupn}}$ (duplicated note): a note predicted whereas it was already being played at the current time being.
- $\mathbf{TSE_{nnof}}$ (no NoteOff): a `NoteOn` token been predicted with no following `NoteOff` token to end it.
- $\mathbf{TSE_{nnon}}$ (no NoteOn): `NoteOff` token predicted whereas this note was not being played.

For each generated token, a rule-based function analyzes its type and value to determine if both are valid, or which type of error was made otherwise. The overall number of errors is normalized by the number of predicted tokens.

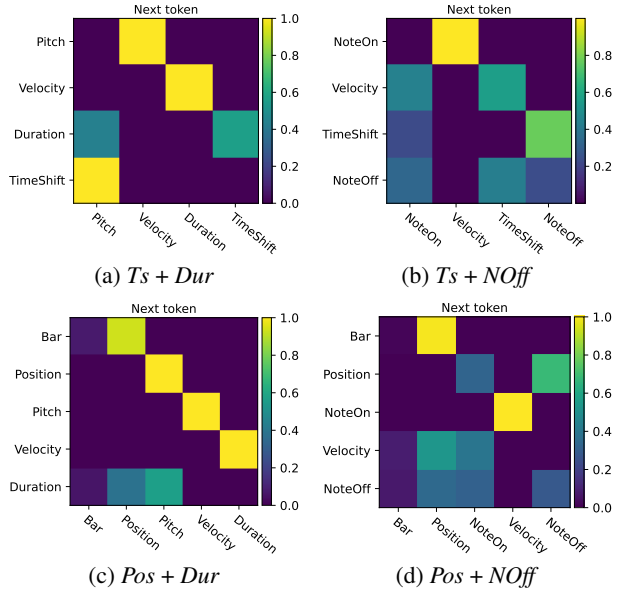
The results are reported in table 2. We first observe that the type error ratios are lower than in other categories. This

[4] Available on the MIDI Manufacturers Association website.



(a) *Ts + Dur* (b) *Ts + NOff* (c) *Pos + Dur* (d) *Pos + NOff* (e) POP909 dataset

Figure 1: Histograms of the note onset positions within bars (top-row), note offset positions within bars (middle-row) and note durations (bottom-row) of the generated notes. There are 32 possible positions within a bar, numerated from 0 (beginning of bar) to 31 (last 32^{th} note). The durations are expressed in beats, ranging from a 32^{th} note to 8 beats.



(a) *Ts + Dur* (b) *Ts + NOff*

(c) *Pos + Dur* (d) *Pos + NOff*

Figure 2: Token type succession heatmaps of the generated results. The vertical axis is a the current token type, the horizontal axis is the next token type following the current one. All rows are normalized.

is excepted since it is less computationally demanding to model the possible next types depending solely on the last one, rather than on the value of the predicted token, for which the validity depends on a the whole previous context.

`Position` tokens bring almost no type errors, but a noticeable proportion of time errors. When decoding tokens to notes, this means that the time may go backward, and resulting in sections of overlapping notes.

Tokenization	$\mathbf{TSE_{type}}\downarrow$	$\mathbf{TSE_{time}}\downarrow$	$\mathbf{TSE_{dupn}}\downarrow$	$\mathbf{TSE_{nnon}}\downarrow$	$\mathbf{TSE_{nnof}}\downarrow$
TS + Dur	$< 10^{-3}$	-	0.014	-	-
TS + NOff	$< 10^{-3}$	-	0.001	0.109	0.040
Pos + Dur	0.002	0.113	0.032	-	-
Pos + NOff	0.002	0.127	0.005	0.095	0.066

Table 2: Prediction error ratios when performing autoregressive generation. - symbol stands for not concerned, and can be interpreted as 0.

Although `Duration` tokens seem to bring slightly more note duplication errors, the use of `NoteOn` and `NoteOff` tokens results in a considerable proportion of note prediction errors. `NoteOff` tokens predicted while the associated note was not being played ($\mathbf{TSE_{nnon}}$) does not have undesirable consequences when decoding tokens to notes, but it pointlessly extends the sequence, reducing the efficiency of the model, and may mislead the next token predictions. Additionaly, `NoteOn` tokens predicted without associated `NoteOff` ($\mathbf{TSE_{nnof}}$) result in notes not properly ended. This error can only be handled by applying a maximum note duration after decoding. Explicit `Duration` tokens allows to specify in advance this information, for both short and long notes. Conversely, with `NoteOff` tokens, the note duration information is implicit and inferred by the combinations of `NoteOn`, `NoteOff` and time tokens. This can be interpreted as an extra effort for the model. Consequently, some uncertainty on the duration accumulates over autoregressive steps during generation. Based on these results, the best tradeoff ensuring good predictions seems to represent time with `TimeShift` tokens and note duration with `Duration` tokens.

In fig. 1 we observe the positions within bars and durations of the generated notes. In all cases, onset positions

are more distributed at the beginning of the bars. This is especially the case with `Bar` and `Position` tokens, for which we may find unexpected rests at the end of bars, when `Bar` tokens are predicted during the generation before that the current bar is completed. The *TS + Dur* combination places note onsets much more on even positions. The probability mass of *TimeShift* tokens (especially for short values) seems to be much higher. However, this is not the case for the *TS + NOff* combination, as `TimeShift` tokens have to be predicted to move the time on odd positions of note offsets. As shown in fig. 2, right after the model is likely to predict a next note, resulting in evenly distributed onset distribution.

Finally, the use of `NoteOff` tokens tends to produce longer note durations, especially when combined with `Position` tokens. In this last case, we can assume that the model might "forget" the notes currently being played, and that it struggles more to model their durations that have to be implicitly deduced from the past `Bar` and `Position` tokens.

Tokenization	Top-20 composers ↑	Top-100 composers ↑	Emotion ↑
TS + Dur	**0.973**	**0.941**	**0.983**
TS + NOff	0.962	0.930	0.962
Pos + Dur	0.969	0.927	0.963
Pos + NOff	0.963	0.925	0.956

Table 3: Accuracy on classification tasks.

5. CLASSIFICATION

For some classification tasks, symbolic music is arguably better suited than audio or piano roll. This is particularly true for classical music feature classification, such as composer [32]. Mono-instrument music with complex melodies and harmonies and no particular audio effect benefit from being represented as discrete for classification and modeling tasks. Given this, it felt important to us to conduct experiments on such task.

We choose to experiment with the GiantMIDI [33] dataset for composer classification and the EMOPIA [34] dataset for emotion classification. The results, as shown in table 3, indicate that there is very little difference between the various tokenization methods. However, the combination of `TimeShift` and `Duration` consistently outperforms the others by one point

The classification task involves modeling the patterns from data that are characteristic to composers or emotions. Here, it seems that the time distance between notes, and their explicit duration play a role in these task, more than note offsets or onset positions. This comes with no surprise for the composer classification task, considering that the data is largely composed of complex music with dense melodies and harmonies, featuring mostly short successive notes. Intuitively, patterns of note successions and chords are more easily distinguishable with explicit durations. With implicit note durations, the overall patterns must be deduced by the combinations of `NoteOn` and `NoteOff` tokens while keeping track of the time.

6. SEQUENCE REPRESENTATION

The last task that we wished to explore is sequence representation. It consists in obtaining a fixed size embedding representation of an input sequence of tokens $p_\theta : \mathbb{V}^L \mapsto \mathbb{R}^d$. Here $\mathbb{V} \subset \mathbb{N}$ denotes the token ids of the vocabulary $\mathcal{V}$, L is the variable input sequence length, and d the size of embeddings. In other words, the model learns to project an input token sequence into a embedding space, thus providing a universal representation. We find this task interesting and well-suited to assess model performances as it directly trains it to model the relationships between tokens within the input sequence and between different representations themselves. While the real-world applications of this task for symbolic music are currently limited, it serves as a useful benchmarking technique for measuring how tokenization impacts the learning of models.

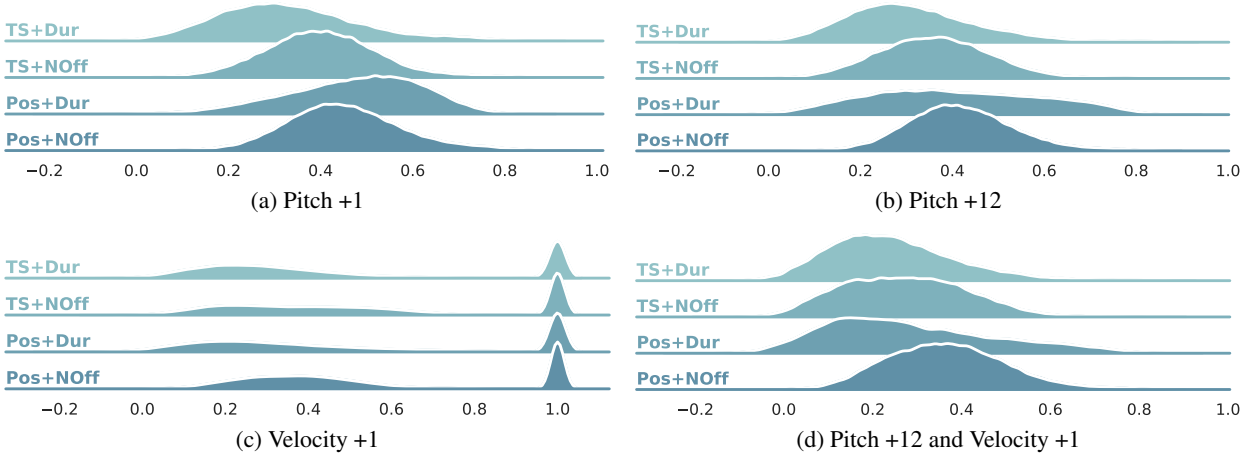
This task has previously been addressed in natural language processing by SentenceBERT [35] or SimCSE [36]. We adopted the approach of the latter, which uses contrastive learning to train the model to learn sequence representations, for which similar inputs have higher cosine similarities. The sequence embedding is obtained by performing a pooling operation on the output hidden states of the model. We decided to use the last hidden state of the `BOS` token position, as it yielded good results with SimCSE [36] [5] . We trained the models with the dropout method: during training, a batch of n sequences $\mathcal{X} = \{\mathbf{x}_i\}_{i=0}^{n}$ is passed twice to the model, but with different dropout masks, resulting in different output sequence embeddings $\mathcal{Z} = \{\mathbf{z}_i\}_{i=0}^{N}$ and $\bar{\mathcal{Z}} = \{\bar{\mathbf{z}}_i\}_{i=0}^{N}$. Although the dropout altered the outputs, most of the input information is still accessible to the model. Hence, we expect pairs of sequence embeddings $(\mathbf{z}_i, \bar{\mathbf{z}}_i)$ to be similar, so having a high cosine similarity. To achieve this objective, we train the model with a loss function defined by the cross-entropy for in-batch pairwise cosine similarities (sim):

$$\ell_i = -\log \frac{e^{\mathrm{sim}(\mathbf{z}_i, \bar{\mathbf{z}}_i)/\tau}}{\sum_{j=1}^{N} e^{\mathrm{sim}(\mathbf{z}_i, \bar{\mathbf{z}}_j)/\tau}} \tag{1}$$

As a result, the model will effectively learn to create similar sequence embeddings for similar inputs, while pushing apart those with dissimilarities. We kept a 0.1 dropout value to train the models, and used the GiantMIDI dataset [33].

Evaluation of sequence representation is intuitively performed by measuring the distances and similarities of pairs of similar sequences. We resort to data augmentation by shifting the pitch and velocity of the sequences in order to get pairs of similar music sequences. The augmented data keeps most of the information of the original data. As such, the models are expected to produce similar embeddings for pairs of original-augmented sequence. Ideally, the cosine similarity should be high, yet not to be equal to 1, as this would indicate that the model fails to capture the differences between the two sequences. The results, presented in fig. 3, indicate that `Position`-based tokenizations per-

[5] SimCSE uses a `CLS` token which is equivalent to `BOS` in our case.



(a) Pitch +1 (b) Pitch +12

(c) Velocity +1 (d) Pitch +12 and Velocity +1

Figure 3: Density plots of cosine similarities between pairs of original and augmented token sequences.

form slightly better. Therefore, it appears that explicit note onset and offset positions information facilitates models to obtains a universal musical representation.

Unlike classification, the contrastive learning objective models the similarities and dissimilarities between examples in the same batch. In this context, note onset and offset positions appear to be helpful for the models to distinguish music.

We also note the contrasting results when augmenting the velocity. Increasing it by one unit, which would be equivalent to playing just a little bit louder, have arguably a very small impact. As a result, the models mostly produces embeddings that are almost identical for the original and the augmented sequences, but also exhibits uncertainty for a notable proportion of samples.

To complement these results, we estimated the isotropy of sets of sequence embeddings. Isotropy measures the uniformity of the variance of a set points in a space. More intuitively, in an isotropic space, the embeddings are evenly distributed. It has been associated with improved performances in natural language tasks [37–39], because embeddings are more equally distant proportionally to the density of their area, and are in turn more distinct and distinguishable. We choose to estimate it with the intrinsic dimension of the sets of embeddings. Intrinsic dimension is the number of dimensions required to represent a set of points. It can be estimated by several manners [40]. We choose Principal Component Analysis (PCA) [41], method of moments (MOM) [42], Two Nearest Neighbors (TwoNN) [43] and FisherS [44]. The results, reported in table 4, show that the embeddings created from the *Pos + Dur* combination tend to occupy more space across the dimension of the model, and are potentially better distributed.

Tokenization	lPCA ↑	MOM ↑	TwoNN ↑	FisherS ↑
TS + Dur	**213**	42.6	34.3	17.5
TS + NOff	161	43.7	32.7	17.5
Pos + Dur	146	39.1	33.1	17.1
Pos + NOff	177	**45.2**	**35.6**	**17.8**

Table 4: Intrinsic dimension of sequence embeddings, as an estimation of isotropy.

7. CONCLUSION

We have discussed the importance of different aspects of symbolic music tokenization, and focused on two major ones: the time and note duration representations. We showed that different tokenization strategies can lead to different model performances due to the explicit information carried by tokens, depending on the task at hand.

Explicitly representing note duration leads to better classification accuracy as it helps the models to capture the melodies and harmonies of a music. Modeling durations, when represented implicitly, adds an extra effort to the model. However, the note offset position information it brings have been found to be more discriminative and effective in our contrastive learning experiment.

For music generation, the time representation plays a significant role, for which the note onset and offsets distributions vary due to the successions of token types. Implicit note durations are less suited for the autoregressive nature of this task, from a prediction error perspective, and sometimes "forgetting" notes being played resulting in higher durations.

We consider this work as a first step into the study of music tokenization for music modeling. We did not experiment with the other music tokenization dimensions, and other important tasks such as music transcription or synthesis for which we could find different results. For transcription, input audio frames are likely to contain the ending or beginning of notes, without being able to identify their onset or offset notes. Explicit `Duration` tokens might confuse models. Furthermore, we believe that more musical reasoning tasks, imposing the model to perform logic deductions to retrieve implicit information from the data, might give more insightful results. Future research will further explore these questions.

8. REFERENCES

[1] J.-P. Briot, G. Hadjeres, and F.-D. Pachet, *Deep Learning Techniques for Music Generation*, ser. Computational Synthesis and Creative Systems. Springer International Publishing, 2020. [Online]. Available: https://www.springer.com/gp/book/9783319701622

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[3] S. Oore, I. Simon, S. Dieleman, D. Eck, and K. Simonyan, "This time with feeling: Learning expressive musical performance," *Neural Computing and Applications*, vol. 32, p. 955–967, 2018. [Online]. Available: https://link.springer.com/article/10.1007/s00521-018-3758-9

[4] Y.-S. Huang and Y.-H. Yang, "Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions," in *Proceedings of the 28th ACM International Conference on Multimedia*, ser. MM '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 1180–1188. [Online]. Available: https://doi.org/10.1145/3394171.3413671

[5] M. Zeng, X. Tan, R. Wang, Z. Ju, T. Qin, and T.-Y. Liu, "MusicBERT: Symbolic music understanding with large-scale pre-training," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, 8 2021, pp. 791–800. [Online]. Available: https://aclanthology.org/2021.findings-acl.70

[6] N. Fradet, J.-P. Briot, F. Chhel, A. El Fallah Seghrouchni, and N. Gutowski, "MidiTok: A python package for MIDI file tokenization," in *Extended Abstracts for the Late-Breaking Demo Session of the 22nd International Society for Music Information Retrieval Conference*, 2021. [Online]. Available: https://github.com/Natooz/MidiTok

[7] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, "The curious case of neural text degeneration," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=rygGQyrFvH

[8] L.-C. Yang and A. Lerch, "On the evaluation of generative models in music," *Neural Comput. Appl.*, vol. 32, no. 9, p. 4773–4784, 5 2020. [Online]. Available: https://doi.org/10.1007/s00521-018-3849-7

[9] C. Helwe, C. Clavel, and F. M. Suchanek, "Reasoning with transformer-based models: Deep learning, but shallow reasoning," in *3rd Conference on Automated Knowledge Base Construction*, 2021. [Online]. Available: https://openreview.net/forum?id=Ozp1WrgtF5_

[10] Q. Zhang, S. Chen, D. Xu, Q. Cao, X. Chen, T. Cohn, and M. Fang, "A survey for efficient open domain question answering," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 14 447–14 465. [Online]. Available: https://aclanthology.org/2023.acl-long.808

[11] Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, and J. Ba, "Large language models are human-level prompt engineers," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=92gvk82DE-

[12] G. Hadjeres, F. Pachet, and F. Nielsen, "DeepBach: a steerable model for Bach chorales generation," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 8 2017, pp. 1362–1371. [Online]. Available: https://proceedings.mlr.press/v70/hadjeres17a.html

[13] B. L. Sturm, J. F. Santos, and I. Korshunova, "Folk music style modelling by recurrent neural networks with long short-term memory units," in *Extended abstracts for the Late-Breaking Demo Session of the 16th International Society for Music Information Retrieval Conference*, 2015. [Online]. Available: https://ismir2015.ismir.net/LBD/LBD13.pdf

[14] W.-Y. Hsiao, J.-Y. Liu, Y.-C. Yeh, and Y.-H. Yang, "Compound word transformer: Learning to compose full-song music over dynamic directed hypergraphs," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 1, pp. 178–186, 5 2021. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/16091

[15] Y. Ren, J. He, X. Tan, T. Qin, Z. Zhao, and T.-Y. Liu, "Popmag: Pop music accompaniment generation," in *Proceedings of the 28th ACM International Conference on Multimedia*. Association for Computing Machinery, 2020, p. 1198–1206. [Online]. Available: https://doi.org/10.1145/3394171.3413721

[16] C. Donahue, H. H. Mao, Y. E. Li, G. W. Cottrell, and J. J. McAuley, "Lakhnes: Improving multi-instrumental music generation with cross-domain pre-training," in *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019, Delft, The Netherlands, November 4-8, 2019*, 2019, pp. 685–692. [Online]. Available: http://archives.ismir.net/ismir2019/paper/000083.pdf

[17] C. Payne, "Musenet," 2019. [Online]. Available: https://openai.com/blog/musenet

[18] J. Liu, Y. Dong, Z. Cheng, X. Zhang, X. Li, F. Yu, and M. Sun, "Symphony generation with permutation invariant language model," in *Proceedings of the 23rd International Society for Music Information Retrieval Conference*. Bengalore, India: ISMIR, Dec. 2022. [Online]. Available: https://arxiv.org/abs/2205.05448

[19] N. Fradet, N. Gutowski, F. Chhel, and J.-P. Briot, "Byte Pair Encoding for symbolic music," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, Dec. 2023. [Online]. Available: https://arxiv.org/abs/2301.11975

[20] M. Kermarec, L. Bigo, and M. Keller, "Improving tokenization expressiveness with pitch intervals," in *Extended Abstracts for the Late-Breaking Demo Session of the 23nd International Society for Music Information Retrieval Conference*, 2022. [Online]. Available: https://ismir2022program.ismir.net/lbd_369.html

[21] G. Hadjeres and L. Crestel, "The piano inpainting application," 2021. [Online]. Available: https://arxiv.org/abs/2107.05944

[22] D. von Rütte, L. Biggio, Y. Kilcher, and T. Hofmann, "FIGARO: Controllable music generation using learned and expert features," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=NyR8OZFHw6i

[23] J. Ens and P. Pasquier, "Mmm : Exploring conditional multi-track music generation with the transformer," 2020. [Online]. Available: https://arxiv.org/abs/2008.06048

[24] H.-W. Dong, K. Chen, S. Dubnov, J. McAuley, and T. Berg-Kirkpatrick, "Multitrack music transformer," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/10094628

[25] O. Khattab and M. Zaharia, "Colbert: Efficient and effective passage search via contextualized late interaction over bert," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 39–48. [Online]. Available: https://doi.org/10.1145/3397271.3401075

[26] J. Ching and y.-h. Yang, "Learning to generate piano music with sustain pedals," in *Extended Abstracts for the Late-Breaking Demo Session of the 22nd International Society for Music Information Retrieval Conference*, 2021. [Online]. Available: https://archives.ismir.net/ismir2021/latebreaking/0000017.pdf

[27] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[28] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-demos.6

[29] P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh, and H. Wu, "Mixed precision training," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=r1gs9JgRZ

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[31] Z. Wang, K. Chen, J. Jiang, Y. Zhang, M. Xu, S. Dai, G. Bin, and G. Xia, "Pop909: A pop-song dataset for music arrangement generation," in *Proceedings of 21st International Conference on Music Information Retrieval, ISMIR*, 2020. [Online]. Available: https://arxiv.org/abs/2008.07142

[32] Q. Kong, K. Choi, and Y. Wang, "Large-scale midi-based composer classification," 2020. [Online]. Available: https://arxiv.org/abs/2010.14805

[33] Q. Kong, B. Li, J. Chen, and Y. Wang, "Giantmidi-piano: A large-scale midi dataset for classical piano music," in *Transactions of the International Society for Music Information Retrieval*, vol. 5, 2021, pp. 87–98. [Online]. Available: https://transactions.ismir.net/articles/10.5334/tismir.80/#

[34] H. Hung, J. Ching, S. Doh, N. Kim, J. Nam, and Y. Yang, "EMOPIA: A multi-modal pop piano dataset for emotion recognition and emotion-based music generation," in *Proceedings of the 22nd International Society for Music Information Retrieval Conference, ISMIR 2021, Online, November 7-12,*

2021, J. H. Lee, A. Lerch, Z. Duan, J. Nam, P. Rao, P. van Kranenburg, and A. Srinivasamurthy, Eds., 2021, pp. 318–325. [Online]. Available: https://archives.ismir.net/ismir2021/paper/000039.pdf

[35] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. [Online]. Available: https://arxiv.org/abs/1908.10084

[36] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple contrastive learning of sentence embeddings," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 6894–6910. [Online]. Available: https://aclanthology.org/2021.emnlp-main.552

[37] T. Wang and P. Isola, "Understanding contrastive representation learning through alignment and uniformity on the hypersphere," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, Jul. 2020, pp. 9929–9939. [Online]. Available: https://proceedings.mlr.press/v119/wang20k.html

[38] D. Biś, M. Podkorytov, and X. Liu, "Too much in common: Shifting of embeddings in transformer language models and its implications," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, Jun. 2021, pp. 5117–5130. [Online]. Available: https://aclanthology.org/2021.naacl-main.403

[39] Y. Liang, R. Cao, J. Zheng, J. Ren, and L. Gao, "Learning to remove: Towards isotropic pre-trained bert embedding," in *Artificial Neural Networks and Machine Learning – ICANN 2021: 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 14–17, 2021, Proceedings, Part V*. Berlin, Heidelberg: Springer-Verlag, 2021, p. 448–459. [Online]. Available: https://doi.org/10.1007/978-3-030-86383-8_36

[40] J. Bac, E. M. Mirkes, A. N. Gorban, I. Tyukin, and A. Zinovyev, "Scikit-dimension: A python package for intrinsic dimension estimation," *Entropy*, vol. 23, no. 10, 2021. [Online]. Available: https://www.mdpi.com/1099-4300/23/10/1368

[41] K. Fukunaga and D. Olsen, "An algorithm for finding intrinsic dimensionality of data," *IEEE Transactions on Computers*, vol. C-20, no. 2, pp. 176–183, 1971. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/1671801

[42] L. Amsaleg, O. Chelly, T. Furon, S. Girard, M. E. Houle, K.-I. Kawarabayashi, and M. Nett, "Extreme-value-theoretic estimation of local intrinsic dimensionality," *Data Mining and Knowledge Discovery*, vol. 32, no. 6, pp. 1768–1805, 11 2018. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01864580

[43] E. Facco, M. d'Errico, A. Rodriguez, and A. Laio, "Estimating the intrinsic dimension of datasets by a minimal neighborhood information," *Scientific Reports*, vol. 7, no. 1, p. 12140, 9 2017. [Online]. Available: https://doi.org/10.1038/s41598-017-11873-y

[44] L. Albergante, J. Bac, and A. Zinovyev, "Estimating the effective dimension of large biological datasets using fisher separability analysis," in *International Joint Conference on Neural Networks (IJCNN)*, 7 2019, pp. 1–8. [Online]. Available: https://arxiv.org/abs/1901.06328